A Comprehensive Study of Knowledge Editing for Large Language Models

Year: 2024 | *Citations: 122* | *Authors: Ningyu Zhang, Yunzhi Yao, Bo Tian*

Abstract

Large Language Models (LLMs) have shown extraordinary capabilities in understanding and generating text that closely mirrors human communication. However, a primary limitation lies in the significant computational demands during training, arising from their extensive parameterization. This challenge is further intensified by the dynamic nature of the world, necessitating frequent updates to LLMs to correct outdated information or integrate new knowledge, thereby ensuring their continued relevance. Note that many applications demand continual model adjustments post-training to address deficiencies or undesirable behaviors. There is an increasing interest in efficient, lightweight methods for on-the-fly model modifications. To this end, recent years have seen a burgeoning in the techniques of knowledge editing for LLMs, which aim to efficiently modify LLMs' behaviors within specific domains while preserving overall performance across various inputs. In this paper, we first define the knowledge editing problem and then provide a comprehensive review of cutting-edge approaches. Drawing inspiration from educational and cognitive research theories, we propose a unified categorization criterion that classifies knowledge editing methods into three groups: resorting to external knowledge, merging knowledge into the model, and editing intrinsic knowledge. Furthermore, we introduce a new benchmark, KnowEdit, for a comprehensive empirical evaluation of representative knowledge editing approaches. Additionally, we provide an in-depth analysis of knowledge location, which can give a deeper understanding of the knowledge structures inherent within LLMs. Finally, we discuss several potential applications of knowledge editing, outlining its broad and impactful implications.